Natural Language Processing to the Rescue?: Extracting "Situational Awareness" Tweets During Mass Emergency

Sudha Verma[1], Sarah Vieweg[2], William J. Corvey[3], Leysia Palen[1], James H. Martin[1], Martha Palmer[3], Aaron Schram[1] & Kenneth M. Anderson[1]

[1]Department of Computer Science, [2]ATLAS Institute, [3]Department of Linguistics
University of Colorado, Boulder, CO 80309
{*sudha.verma, sarah.vieweg, william.corvey, leysia.palen, james.martin, martha.palmer, aaron.schram, kenneth.m.anderson*}@colorado.edu

Abstract

In times of mass emergency, vast amounts of data are generated via computer-mediated communication (CMC) that are difficult to manually cull and organize into a coherent picture. Yet valuable information is broadcast, and can provide useful insight into time- and safety-critical situations if captured and analyzed properly and rapidly. We describe an approach for automatically identifying messages communicated via Twitter that contribute to situational awareness, and explain why it is beneficial for those seeking information during mass emergencies.

We collected Twitter messages from four different crisis events of varying nature and magnitude and built a classifier to automatically detect messages that may contribute to situational awareness, utilizing a combination of hand-annotated and automatically-extracted linguistic features. Our system was able to achieve over 80% accuracy on categorizing tweets that contribute to situational awareness. Additionally, we show that a classifier developed for a specific emergency event performs well on similar events. The results are promising, and have the potential to aid the general public in culling and analyzing information communicated during times of mass emergency.

Introduction

During mass emergency, time-sensitive requirements arise; people may need food, shelter, and medical care, among other essentials. Additionally, the affected as well as helping populations need information; the emergency periods of warning, impact and response are marked by intensive information search (Starbird et al., 2010). The pervasiveness of information and communication technology (ICT), including social media like networking sites and microblogging services, has opened the proverbial floodgates of information dissemination.

A recent surge in research on computer-mediated communication during mass emergency situations looks at content in blogs, social media sites and microblogging services (Mendoza et al., 2010; Palen et al., 2009; Qu et al., 2009; Starbird et al., 2010; Vieweg et al., 2010). This work shows that members of the public turn to social networking sites, microblogging services and similar technologies to understand and communicate during emergency situations. We extend this research by focusing on Twitter communications (tweets) generated during mass emergency, and show how Natural Language Processing (NLP) techniques contribute to the task of sifting through massive datasets when time is at a premium and safety of people and property is in question.

So much information is now broadcast during mass emergencies that it is infeasible for humans to effectively find it, much less organize, make sense of, and act on it. To locate useful information, computational methods must be developed and implemented to augment human efforts at information comprehension and integration.

The popular microblogging service Twitter serves as an outlet for many to offer and receive useful information; it provides a way for those experiencing a mass emergency to gather more, or different, information than they may be able to using mainstream media and other traditional forms of information dissemination. This access provides affected populations with the possibility to make more informed decisions. The challenge, however, is in locating the *right* information. In addition to broadcasting valuable, actionable information via Twitter during mass emergency, many also send general information that is void of helpful details, or communicate empathetic, supportive messages that lack tactical information.

Tweets that include tactical, actionable information contribute to *situational awareness*; such tweets include content that demonstrates an awareness of the scope of the crisis as well as specific details about the situation. We offer an approach for automatically locating information that has the potential to contribute to situational awareness in the multitude of tweets broadcast during mass emergency. Our overarching goal is to help affected populations cull and analyze pertinent information

communicated via computer-mediated communication. Our assumption is that immediate, dynamic culling of tweets with information pertaining to situational awareness could be used to inform and update applications aimed at helping members of the public, formal response agencies, aid organizations and concerned outsiders understand and act accordingly during mass emergencies.

Using NLP and machine learning (ML) techniques, we develop a suite of classifiers to differentiate tweets across several dimensions: subjectivity, personal or impersonal style, and linguistic register (formal or informal style). Based on initial analyses of tweet content, we posit that tweets that contribute to situational awareness are likely to be written in a style that is objective, impersonal, and formal; therefore, the identification of subjectivity, personal style and formal register could provide useful features for extracting tweets that contain tactical information. To explore this hypothesis, we study four mass emergency events: the North American Red River floods of 2009 and 2010, the 2009 Oklahoma grassfires, and the 2010 Haiti earthquake.

The Disaster Events

The Red River floods occurred in March-April 2009 in North Dakota, Minnesota, and Manitoba. The river reached a record high level, but damage remained under control due to mitigation efforts (Starbird et al., 2010). During the same time period as the 2009 Red River Flood, the Oklahoma grass fires took place on April 9-10 2009. These fires caused over 60 injuries and burned more than 100,000 acres in central and southern Oklahoma (Vieweg et al., 2010). Less than a year later, the Haiti earthquake on January 12, 2010 resulted in catastrophic damage, injury and death. The earthquake affected an entire country and spawned an international response (Romero & Lacey 2010). Later in February-March 2010, flooding of the Red River was again a concern. Area residents and local and state government organizations took precautions; minimal damage occurred (FEMA, 2010.)

Theoretical Background

When faced with mass emergency, people go through a process of "situational analysis" (Matheus et al., 2003) to ascertain "situational awareness" (Sarter and Woods, 1991), a state of understanding the "big picture" during times of danger. However, having knowledge of an emergency event at a high level—knowing that there is threat of fire or impending flood—is not enough. People require knowledge of fire location or flood levels, among other details. This specific information is sometimes communicated via Twitter, but finding it can be cumbersome and time-consuming.

We situate this research at the intersection of crisis informatics, the theory and practice of monitoring information flow in computer-mediated communication during and after crisis events (Palen et al., 2009), and a growing body of research on microblogging, specifically regarding information extraction and text-based forecasting. Ultimately, to learn about the human practices that motivate microblogging, one must first characterize the structure and content of microblogs and social media at various levels of granularity. We draw attention to several recent studies on weblogs and social media that are representative of three critical facets of the present research: modeling the topic (semantic content) of a tweet or tweet stream; modeling sentiment within the tweet; and modeling the amount and character of subjectivity represented in a tweet. Our classification features are designed to not only characterize observed tweet behavior, but to form a model that predicts the information value of future tweets with similar characteristics based on the information values of tweets in the training data. Because our eventual goal is to operate in near real-time, classification must occur quickly, over either static or dynamic datasets. Thus, while the current study utilizes static training data, we situate the paper within a body of literature focused on time-series as well as static representations of data.

Of our current classification features, two have attained significant attention in the microblogging literature: subjectivity and sentiment. Subjectivity classification is presupposed in sentiment analysis, as only subjective expressions contain sentiment-bearing constituents. However, subjectivity classification is useful in its own right, for discovery of sarcasm in text (Tsur et al., 2010), or as in our data, for assessing the amount of emotion or opinion content a user expresses in tweets, and the utility of subjective tweets in information extraction or assessment of situational awareness. Much research has concerned the discovery of sentiment polarity in text (e.g., Bruce and Wiebe, 1999), with some recent research correlating sentiment classification with real-world behavior (Gilbert and Karahalios 2010; Tumasjan et al., 2010). Our motivation for sentiment analysis in the crisis informatics domain is to assess the polarity of the emotional content expressed to more accurately assess conditions on the ground.

Finally, the input for our classification scheme consists of tweets that have been hand-labeled as *on-topic*. On-topic tweets mention an event in some way (providing tactical information, describing personal reactions, or acknowledging that an event is occurring/has occurred.) Previous research on crisis data (Vieweg et al., 2010) has indicated that a simple keyword or hashtag-search

approach to tweet collection cannot guarantee extraction of only relevant tweets. Therefore, automatic topic modeling of tweet content is an ongoing interest (Ramage et al., 2010).

We examine additional characteristics of discourse such as how "personal" tweet messages are (i.e. if they indicate shared background or understanding, and if the sender of the tweet injects him/herself into the message in some way) as well as linguistic register (whether they are written in a formal or informal style). We now proceed with a more thorough discussion of classification features and our human annotation process.

Tweet Content Analysis

NLP classifiers require "training" data in the form of annotated, or coded text, which they use to "learn" how to distinguish between different types of discourse. Discourse analysis performed on tweets from our four datasets provided the basis for the qualitative codes we used to annotate tweets.

Dataset Details

For this preliminary study we collected a random sample of roughly 500 tweets broadcast during each of the emergency events that contained one of our keywords. 500 tweets were selected as the target because they are both a manageable set for human annotation and provide sufficient training data for ML classification. The OK fire (OK; 527 tweets) and Red River flood (RR09; 453 tweets) were previously collected and sampled (Starbird et al., 2010; Vieweg et al., 2010). We began the present work using these datasets, which contained on-topic tweets authored by individual Twitterers (i.e. not organizations or businesses) who were located in or near an area affected by each event.

In 2010, we collected tweets sent during the 2010 Red River flood (RR10). The RR10 dataset was collected during a 20-day period, from March 15-April 3 using the following search terms: *fmflood, flood10, red river, redriver, ccflood,* and *fargoflood.* One RR10 tweet from the sample of 500 random tweets was off-topic, resulting in 499 tweets for the purposes of this research. The Haiti dataset is much broader. It consists of tweets collected over a 23-day period representing the day prior to the earthquake until one week after the Haitian government officially ceased search and rescue operations. The entire Haiti dataset consists of tweets that include any of the following search terms: *haiti, earthquake, quake, shaking, tsunami, ouest, Port-au-Prince, tremblement* and *tremblement de terre.* 14 Haiti tweets of the sample of 500 random tweets were off-topic, leaving us with 486 tweets. The four datasets provide us with a variety of tweets broadcast during three different types of mass emergency. Our final dataset contained 1,965 tweets, all of which were hand-annotated. In addition, the Red River and Oklahoma datasets represent "local" communications, while the Haiti dataset is more representative of "international" communications.

Qualitative Coding

Preliminary analysis of tweet content suggested a correlation between tweets expressed in an objective manner and those that contribute to situational awareness. Such tweets communicate, for example, the location of hazards, or the state of recovery efforts. On the other hand, subjective tweets seem to correlate with tweets that *do not* contribute to situational awareness. These tweets communicate offers of sympathy, or requests for support or prayer. Our subsequent analysis caused us to consider additional linguistic styles in our examination of tweets that contribute to situational awareness: *linguistic register*, and whether tweets were written in a *personal* or *impersonal manner*. Though all types of communications are part of the collective behavior we observe among Twitter users in times of mass emergency, and we do not claim that tweets contributing to situational awareness are more important than those that do not; it might be that tweets with certain characteristics tend to contribute to an action-oriented segment of the information space. Such tweets inform decision-making processes, while other tweets contribute to an emotion-oriented segment of the information space.

Two annotators each independently coded a set of tweets from each dataset for four different qualities: whether tweets communicate situational awareness information, if they are objective or subjective, if they are written in a formal or informal style, and if they are written from a personal or impersonal standpoint. An expert annotator then adjudicated these results to form Gold Standard training data. An explanation of each of these categories follows, along with examples from each dataset in this order: OK Fire, RR09, RR10, Haiti.

Situational Awareness Tweet Content

Tweets were coded for whether they contribute to situational awareness based solely on tweet text (we did not analyze links to websites and photos). The tweets shown below contain information that people can use to gauge situations and inform decision-making processes; they were coded as contributing to situational awareness:

> Niece and her dad are being evac'd in MWC, fire is headed towards SIL and her boyfriend in Harrah

> Country residents outside of Fargo are surrounded by flood waters. Some R being rescued.

> The Red River at Fargo ND is at 33.15 ft which is 15.15 ft above flood stage #flood10 #fargoflood

Conversely, tweets that mention an event, but do not provide actionable information, were coded as *not* contributing to situational awareness. The following tweets were coded as such—they mention the events but do not contain information that affected populations can use to make timely, informed decisions:

> Tweeps please pray for the families and firefighters battling these crazy fires in Oklahoma

> Nothing upsets me more than listening to my Dad's voice. He was born in a drought, and now watches his hometown flood

> Got a flood photo of mine published in the local paper today. Time for a new career?

Objective Tweet Content

The following tweets were coded as "objective"—they provide purely factual information and do not express opinions or include sentiment-bearing words:

> OHP is responding to Cater County to assist with RD closures and evacuations due to fires in and near Tutums. Approx 20 homes in danger

> Here is a list of the earthquake survivors. You may find update from Hotel Montana: http://bit.ly/7o7mUk

Subjective Tweet Content

Subjective tweets *do* include opinions and/or convey sentiment, as seen below:

> so proud to be from Oklahoma. The outpouring of support for those devestated by the fires is amazing

> I think they're being a bit over dramatic about the flooding. Trying to make us fearful but like always it wouldn't be close to bad at all

Register

An additional aspect of language use involves linguistic *register*. According to Michael Halliday, "register is determined by what is taking place, who is taking part, and what part the language is playing" (Halliday, 1978, 23). Register is frequently viewed in terms of formality—from very formal and rigid on one end, to very casual on the other. In the case of Twitter use during mass emergency situations, we note that some users employ a formal register, and others an informal register.

Formal tweets are those that are grammatically coherent and express complete thoughts, such as these:

> Staging for fires in Elk City Area is currently at Elk City FD

> 4 more GCC volunteers have deployed to Fargo to help with the feeding and sheltering efforts along the Red River.

Conversely, informal tweets tend to be fragmented, lacking context, include slang and/or several abbreviations, and have many grammatical mistakes:

> @user6 yup homes in Tonua's neighborhood are on fire

> landed in fargo today...locals say red river will crest at 43 feet...worse then 97 flood

Personal/Impersonal Tweet Content

The final quality we focused on was whether tweets were expressed from a personal standpoint or not. Personal tweets indicate that the Twitterer is injecting him or herself into the situation in some way, such as these examples:

> Our best hopes and wishes go out to the folks in Manitoba. As the Red River is about to crest.

> Thinking of my friends in Fargo, and my time spent there. I will keep you all in my prayers. #FMflood

Conversely, impersonal tweets display a sense of emotional distance from the event by the tweet author:

> Canadian and Oklahoma Counties under RED FLAG FIRE WARNING until 10 p.m. This warning means conditions are ripe for wildfire outbreaks.

> The Red River at Fargo is 40.76 feet. 22.76 feet above flood stage. 0.66 feet above 1897 record. #flood09 #fargoflood

Personal/Impersonal coding can be seen as a corollary to subjectivity analysis. Tweets that demonstrate personal characteristics are often also subjective, but it is possible for a tweet to contain subjective content without being expressed from a personal point of view, and it is equally possible for a tweet to be stated from a personal perspective without containing subjective content, as in the following tweet published during the Oklahoma Fires:

> @User13 The fires are away from our area, but I know people who live in some of the areas where houses burned. One works with my SIL

Qualitative Coding Results

In summary, the categories utilized in qualitative coding are meant to satisfy three criteria. First, annotation categories should be intuitive to annotators with minimal training. We expect variation in agreement rates across the

datasets, and we interpret excellent agreement as an indication of a strong delineation between a given coding dyad on a specific disaster and poor agreement as an indication of a weak delineation of a given coding dyad. Annotation training materials consist of token examples and short description of each coding dyad. In addition, annotation categories should be domain-independent. Categories remain consistent across all datasets, and are crisis-specific. Finally, for ease of implementation, annotation categories are labeled as binary features. Inter-tagger agreement (comparing one annotator against the other; ITA) and Kappa (κ) statistics for annotator agreement across categories are shown in Table 1 below; Kappa (Cohen, 1960) is the ITA adjusted for the probability of agreement by chance. Because all codes are binary, the probability of chance agreement is .5.

	Objectivity		**Register**		**Pers./Imp.**		**SA**	
	ITA	κ	ITA	κ	ITA	κ	ITA	κ
RR09	.93	.86	.79	.57	.92	.84	.80	.60
RR10	.92	.83	.90	.80	.95	.89	.91	.81
Haiti	.89	.78	.86	.72	.86	.72	.98	.96
OK Fire	.97	.94	.90	.80	.93	.86	.91	.81

Table 1: ITA and Kappa statistics for inter-rater agreement across all datasets for each annotation category.

Objectivity and Personal/Impersonal codes show consistently high agreement across datasets, which reflects the intuitive nature of the task. Agreement on Register is slightly lower, possibly owing to the fact that this is a more abstract category. Additionally, the agreement values reflect interesting variations among the datasets. For instance, when dealing with a large-scale, catastrophic crisis such as the Haiti earthquake, separating tweets that mention support and prayer from Situational Awareness tweets that offer tactical information seems more straightforward. On the other hand, Register and Situational Awareness may be more difficult to identify in RR09, but it is not immediately obvious why this task is less challenging in RR10. Below we discuss the classification process and results.

Classification

For machine learning, we are interested in answering the following questions:

1) Can tweets demonstrating situational awareness (SA tweets) be automatically identified using machine learning techniques?
2) How well can we categorize other linguistic traits, such as subjectivity, which seem to co-occur with SA tweets? Can we use existing tools to classify well-studied concepts, such as subjectivity?
3) Do the added linguistic features improve classification of SA tweets over standard-derivable features?

Because automatic identification of situational awareness is a novel task, we did not have any *a priori* assumptions about specific linguistic features for classification. Therefore, as a first step, we implemented an SA classifier based on low-level features that can be easily derived from tweets. In parallel, we also implemented three simple classifiers to determine linguistic features—the predicted tags were then used as a feature for the SA Classifier and results evaluated against the hand-annotated tags. To determine the overall scalability of the classifier, we also tested each classifier trained on one type of crisis-event with data from a different event.

Classification Methods

We experimented with two standard machine-learning methods for classification, Naïve Bayes (NB) and Maximum Entropy (MaxEnt). Both have been known to do well for the similar tasks of sentiment analysis and text classification. MaxEnt yielded better results overall. We used Mallet's implementation of the algorithms (McCallum, 2002) and made a few enhancements to Mallet classes to process our data formats.

Data Distribution

A subset of tweets from the four events was annotated at the tweet level to indicate four different traits. These included whether the tweet demonstrated situational awareness, was personal in nature, was written in a formal register and if it was subjective. Table 2 shows the characteristics of hand-annotated features. It shows distribution of objective, impersonal and formal tweets among situational aware tweets (SA) and also not situational aware tweets (~SA).

	Objective		**Impersonal**		**Formal**	
	SA	~SA	SA	~SA	SA	~SA
RR09	87%	56%	88%	53%	81%	46%
RR10	90%	51%	90%	48%	89%	48%
Haiti	83%	50%	78%	54%	76%	41%
OK Fire	88%	80%	64%	34%	77%	51%
Uniform	87%	53%	81%	47%	85%	44%

Table 2: Training data distribution of hand-tagged features among the tweets that are situational aware (SA) and the tweets that are not SA (~SA).

Notably, data collected from localized disasters using a keyword search yield a high number of tweets that include SA content. This is not surprising, as disasters such as the Red River floods and Oklahoma fires are localized events, and Twitter communications reflect this. However,

catastrophic disasters attaining global attention—such as the Haiti earthquake—generate significantly more noise and traffic on Twitter, with a smaller subset of tweets demonstrating situational awareness. As a point of comparison, the average incoming tweet rate for Red River floods generated a few thousand tweets per day, whereas the Haiti earthquake generated up to 800K tweets per day.

We also note that a large proportion of tweets demonstrating situational awareness are objective, formal and impersonal in nature, confirming our belief that these features could have the potential to improve SA classification. On the other hand, among the messages that do not include situational awareness information, distribution of these features is more uniform.

Feature Extraction

Tweets tend to include symbols and artifacts that can confuse the classification process. We replace URLs and Twitter-specific symbols, i.e. "RT", "@username" and the hash symbol (#) with a unique symbol. Standard stop words (e.g., "a," "the,") were also removed. We then tokenized the tweets and used the words and their frequency as features for classification. We also used the Stanford part-of-speech (POS) tagger (Toutanova et al., 2003) to tag the tweets and use the POS tags as features.

The list of features to the SA Classifier includes:

1) Unigrams and their raw frequency.
2) Bigrams and their raw frequency.
3) Part-of-speech tags.
4) Subjectivity of tweets (Objective/Subjective).
    a. Hand-annotated
    b. Predicted by the subjectivity classifier.
    c. Derived by the OpinionFinder (Riloff and Wiebe, 2003; Wiebe and Riloff, 2005).
5) Register of tweet (Formal/Informal)
    a. Hand-annotated
    b. Predicted by register classifier.
6) Tone of tweet (Personal/Impersonal)
    a. Hand-annotated
    b. Predicted by classifier.

Experiments and Evaluation

We implemented classifiers to predict subjectivity, register and tone of the tweet. We evaluated the performance of the SA classifier by introducing one feature at a time on datasets from four different events. To get a more uniform distribution of SA and non-SA tweets, we also created a fifth dataset by randomly selecting 500 situational and 500 non-situational awareness tweets from all events. Performance was evaluated by taking the mean accuracy over 10 fold cross-validation and splitting the data 90% for training and 10% for testing. For our baseline, we used words (unigrams) and their frequency.

Results

Results are detailed in Table 3. We now discuss the relative effectiveness of individual classification features.

Words and Frequency (unigrams/bigrams)

We find high baseline accuracy for events using only word (unigram) and raw frequency as a feature. We hypothesize this is due to emergency events using specific vocabulary to convey situational awareness. For example, in the case of floods, we see that SA tweets include words such as: "depth," "water," "level," and "crested." SA tweets for the OKFire include: "grassfire," "wildfire," and "burn." In the uniform data split, we see higher frequency of nonspecific, but emergency-related words: "safe," "evacuate," and "rescue." These words are infrequent in tweets that do not include information indicating situational awareness.

The addition of bigrams to the model did not yield significant improvement over using unigrams and frequencies alone. This falls in line with similar work done in sentiment analysis (e.g., Pang and Lee, 2002).

Features used	RR09		RR10		Haiti		OKFire		Uniform	
	NB	ME	NB	ME	NB	ME	NB	ME	NB	ME
Unigrams (Baseline)	74.0	82.2	74.4	**89.0**	81.4	81.4	84.3	83.4	79.8	79.3
Unigrams, Bigrams	73.2	81.1	71.0	88.7	79.8	80.2	83.2	82.4	80.3	79.8
Unigrams, POS	75.3	79.1	80.0	87.8	83.3	83.9	83.2	82.6	80.8	83.3
Objective (annotated), POS, unigrams	77.7	84.0	82.2	88.7	82.8	86.5	83.8	84.1	81.0	83.0
Objective (predicted), POS, unigrams	76.5	**84.8**	81.0	88.4	82.2	85.3	79.4	84.7	79.2	82.3
Objective (OpinionFinder), POS, unigrams	71.5	82.4	76.4	85.8	82.0	84.4	80.2	82.6	82.3	81.9
Formal (annotated), POS, unigrams	76.4	82.7	82.2	88.8	84.2	86.2	82.3	**87.9**	80.0	81.8
Formal (predicted), POS, unigrams	75.2	82.1	80.8	87.4	83.3	84.7	79.8	84.4	79.7	81.3
Impersonal (annotated), POS, unigrams	78.2	84.0	82.2	89.0	84.5	86.3	83.0	85.5	83.8	83.5
Impersonal (predicted), POS, unigrams	75.9	83.9	81.4	87.1	83.1	83.7	80.2	85.1	79.8	81.5
All features (predicted), POS, unigrams	76.1	84.1	80.2	88.6	83.5	**88.8**	81.0	87.1	80.2	**84.5**

Table 3: Average 10-fold cross validation accuracies for SA classifier. Best performance for each dataset is in boldface.

Part-of-Speech Tags

We find that for most datasets, part-of-speech tags improve overall classification accuracy. In addition to basic POS tagging, we took a particular interest in tagging only adjectives, since adjectives tend to indicate subjectivity. This did not, however, improve the SA accuracy over POS tags. To select messages written in a personal tone, we also tried using POS tags only if they were personal, possessive pronouns. This also did not help with overall accuracy, implying that the contextual information of the tweet is what helps with SA categorization, rather than the presence or absence of one trait.

Objectivity

All tweets annotated for situational awareness were also annotated for objectivity. We then trained a supervised MaxEnt classifier to classify subjective and objective tweets. We used unigrams and part-of-speech tags as input features for the classifier. POS improved accuracy of the subjective classification. On average, the objective classifier achieved an accuracy of 86.2%.

We evaluated SA classifiers using both gold standard tags and predicted tags, to see if accuracy improves when objectivity of a tweet is available. We see the highest increase in accuracy of the SA classifier when the gold standard tag for objectivity is added as a feature. Predicted objectivity tags also show good improvement in many cases when selecting SA tweets.

We also evaluated the use of OpinionFinder for objectivity/subjectivity analysis (Riloff and Wiebe, 2003; Wiebe and Riloff, 2005). OpinionFinder uses lexical and contextual information to determine a subjectivity tag, and thus does not require us to annotate for subjectivity and train a separate classifier. However, OpinionFinder did not perform as well as our objectivity classifier. This is not entirely surprising since OpinionFinder was not developed with Twitter-like data in mind, and indicates the value of domain-specific subjectivity classification.

Formal Register and Impersonal Tone

Categorizing tweets as Formal/Informal and Personal/Impersonal tended to improve situational awareness classification. Similar to subjectivity, a MaxEnt classifier was trained on the hand-annotated data for predicting register of a tweet. We used the same feature set as the subjectivity classifier. Register of a tweet proved harder to predict than subjectivity, giving us an average accuracy of 78.4%. The classifier to categorize tweets for Personal and Impersonal style resulted in 85.7% average accuracy. For both register and personal/impersonal style, using the predicted tags yielded similar accuracies as using the hand-annotated tags.

Summary of Feature and Accuracy Results

There is some overlap between objective, impersonal and formal traits. For example, objective tweets tend to be written formally and impersonally. Using any one of the features improved the accuracy for classifying tweets with SA content, and using all the features together helped improve overall performance even more. The amount the linguistic features helped varied from dataset to dataset. In a uniform distribution, addition of the linguistic features reduced error rate by 11% over using only part-of-speech tags as features. However, in real crisis events, we have not encountered uniform distributions. For disasters that are not large-scale, keyword searches for the events yield proportionally more SA tweets. For example, in both the Oklahoma fire and 2009 and 2010 Red River floods, more than 75% of the tweets included information indicating situational awareness. Catastrophes such as the Haiti earthquake have a greater signal to noise ratio. In both cases, additional linguistic features greatly reduced the error rate and improved overall accuracy of the classifier, well beyond part-of-speech tags.

Scalability Experiments and Evaluation

After examining features for improvement in classification accuracy, we evaluated how well SA classification performed across events. In a deployed system, we envision a set of classifiers trained on previous events to be used to locate SA tweets for a new event of similar type.

We used a classifier trained on data from previous events to classify data from new events, using those features that gave the best results in the training data. For test data, we considered the text of the new tweets. The MaxEnt classifier performs better on all datasets. The results of cross-event experiments are in Table 4.

	Test Set			
Train Set	**Haiti**	**RR09**	**RR10**	**OKFire**
Haiti		69.3	73.9	56.4
RR09	50.8		**83.8**	65.1
RR10	52.3	**80.1**		69.3
OKFire	29.0	81.4	87.0	

Table 4: Average 10-fold cross validation accuracies on cross-event classification using MaxEnt. Best performance in bold.

Generally, classification of a new event using a classifier trained on previous events was more challenging, but performance is better for events of similar type. For instance, when a tweet was labeled using a classifier trained on a similar disaster, results were promising; a classifier trained on RedRiver09 gave an accuracy of 84% on RedRiver10. The disaster in Haiti, however, differed from others in scope of the media coverage, the type of disaster and the extent of damage, injury and death. These

factors most likely contributed to the poor performance when using the Haiti dataset to classify other events and when using other events data to classify it. These aspects of classification will be worth exploring in our future work.

Discussion

We demonstrate that a classifier based on low-level linguistic features performs well at identifying tweets that contribute to situational awareness. Further, we show that linguistically-motivated features including subjectivity, personal/impersonal style, and register substantially improve system performance. These results suggest that identifying key features of user behavior can aid in predicting whether an individual tweet will contain tactical information. In demonstrating a link between SA and other markable characteristics of Twitter communication, we not only enrich our classification model, we also enhance our perspective of the space of information disseminated during mass emergency.

In future work, we will explore features that help lower false positive and negative rates. False positives are of greater concern, since they represent noise that could be misleading. To address them, we will explore incorporating user feedback; increasing the weight of tweets that have been retweeted; and the effects of using different limits in our machine learning algorithms. False negatives represent SA messages overlooked by the classifier. Because Twitter has a high degree of redundancy, it is less likely that all tweets that represent the same information and are written in different styles will be misclassified.

To measure classifier accuracy, we tested a sample of manually annotated tweets. As a deployed system, it will continuously classify incoming tweets based on models built on data from previous similar events. We will expand our corpora both in size and event type. We will also explore using semi-supervised approaches to increase our coverage and depth to determine if the system scales to different quantities of computer mediated communication.

Acknowledgements

We gratefully acknowledge sponsorship from the US National Science Foundation through grants IIS-0546315 and IIS-0910586. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do reflect the views of the National Science Foundation.

References

Bruce, R.F. & J.M. Wiebe. 1999. Recognizing subjectivity: a case study in manual tagging.*Natural Language Engineering 5:2,*1-16.

Cohen, J. 1960. A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement* 20: 37-46. FEMA: National Situation Update: Wednesday, March 24, 2010. http://www.fema.gov/emergency/reports/2010/nat032410.shtm

Gilbert, E. & K. Karahalios. 2010. Widespread Worry and the Stockmarket. *Proc. ICWSM 2010*, 58-65.

Halliday, M.A.K. 1978. *Language as social semiotic*. Baltimore, MD: University Park Press.

Matheus, C.J., Kokar, M.M. & K. Baclawski. 2003. A Core Ontology for Situational Awareness. *Proc. Information Fusion*.

McCallum, A.K. 2002. "MALLET: A Machine Learning for Language Toolkit." http://mallet.cs.umass.edu.

Mendoza, M., Poblete, B. & C. Castillo. 2010. Twitter Under Crisis: Can we trust what we RT? KDD Social Media Analytics Workshop, *Proc. SOMA 2010*.

Palen, L., Vieweg, S., Liu, S.B. & A.L. Hughes. 2009. Crisis in a Networked World: Features of Computer-Mediated Communication in the April 16, 2007 Virginia Tech Event. *Social Science Computer Review 27:4,* 1-14.

Pang, B., Lee, L., Lillian & S. Vaithyanathan. 2002. Thumbs up? Sentiment classification using machine learning techniques. In *EMNLP*, 79–86.

Qu, Y., P.F. Wu, & X. Wang. 2009. Online Community Response to Major Disaster: A Study of Tianya Forum in the 2008 Sichuan Earthquake. *Proc. HICSS 2009.*

Ramage, D., Dumais, S., and D. Liebling. 2010. Characterizing Microblogs with Topic Models. *Proc. ICWSM 2010*.

Riloff, E. and J. Wiebe. 2003. Learning Extraction Patterns for Subjective Expressions. *Proc. Conference on Empirical Methods in Natural Language Processing (EMNLP-03). ACL SIGDAT*, 105-112.

Romero, S. & M. Lacey. Jan. 12, 2010. New York Times. Fierce Quake Devastates HaitianCapital, http://www.nytimes.com/2010/01/13/world/americas/13haiti.html

Sarter, N.B. & D.D. Woods. 1991. Situation Awareness: A Critical but Ill-Defined Phenomenon. *The International Journal of Aviation Psychology 1:1*, 45-57.

Starbird, K., Palen, L, Hughes, A.L. & S. Vieweg. 2010. Chatter on the Red: What Hazards Threat Reveals about the Social Life of Microblogged Information. *Proc. CSCW 2010,* 241-250.

Toutanova, K., Klein, D., Manning, C., & Y. Singer. 2003. Feature-Rich Part-of-Speech Tagging with a Cyclic Dependency Network. *Proc. HLT-NAACL 2003*, 252-259.

Tsur, O. Davidov, D. & A. Rappoport. 2010. ICWSM—A Great Catchy Name: Semi-Supervised Recognition of Sarcastic Sentences in Online Product Reviews. *Proc. ICWSM 2010*, 162-169.

Tumasjan, A., Sprenger, T.O., Sandner, P.G. & I.M. Welpe. 2010. Predicting Elections with Twitter: What 140 Characters Reveal about Political Sentiment. *Proc. ICWSM 2010*, 178-185.

Vieweg, S., Hughes, A.L., Starbird, K. & L. Palen. 2010. Microblogging During Two Natural Hazards Events: What Twitter May Contribute to Situational Awareness. 2010. *Proc. CHI 2010*, 1079-1088.

Wiebe, J.M. and E. Riloff. 2005. Creating subjective and objective sentence classifiers from unannotated texts. *Proc. CICLing-2005*.